**Human Activity Recognition**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Project Mentor: Dr. Namitha K**

**Project Lead: Girish S**

**Team Members**

| Girish S | AM.EN.U4AIE22044 |
| --- | --- |
| Anuvind MP | AM.EN.U4AIE22010 |
| R S Harish Kumar | AM.EN.U4AIE22042 |
| Harishankar Binu Nair | AM.EN.U4AIE22023 |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

| **File** | **Description** | **Status** |
| --- | --- | --- |
| FileAutism |  | In progress |
|  |  | Not started |
| FileParkinsons |  | In progress |
| FileStress and Anxiety |  | In progress |
| File |  | Not started |

**Vision-based activity recognition in children with autism-related behaviors**

-> ASD diagnosis, preprocessed data, collected data, human detection, background removal, temporal convolutional models, light weight, Inflated 3D convnet and Multi Stage Temporal Convolutional Network, Weighted F1 score = 0.83, 3 activities, ESNet backbone weighted F1 score =0.71, uncontrolled Environment

-> [“ASD is a neurodevelopmental disorder characterized by a set of social communication deficits, self-harm, or persistent repetition of actions”](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibABB7B79D860AFD1314264C986AF4606Cs1)

-> “ASD related biomarkers such as those determined through functional magnetic resonance imaging (fMRI). [Facial expressions](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibD2F6BC878DD136CE1116D13D2DF5BCA9s1), [eye gaze, and motor control/movement patterns](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibF685F61EDB75CC4742D43DD621AEA962s1)”

-> dataset from SSBD from [rajagopalan et al](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibA1DD584B1FB446E21CB1B438E6012CB9s1).

-> 75 videos from yt, rajagopalan et al provides no other novel methods other than the “new Self-Stimulatory Behavior Dataset (SSBD) for ASD”

-> additionally added videos from online platforms => not mentioned sources

-> Feature extraction (CNN model) => action classification (TCN)

-> dataset transformed into 168 short clips after removing noisy data and adding new data

-> [168 clip dataset](https://github.com/Samwei1/autism-related-behavior/blob/main/url_list.pdf)

-> “used compressed feature representation for real time throughput”?

-> two crucial steps:

i) a feature extractor that extracts frame-wise action-related semantic features

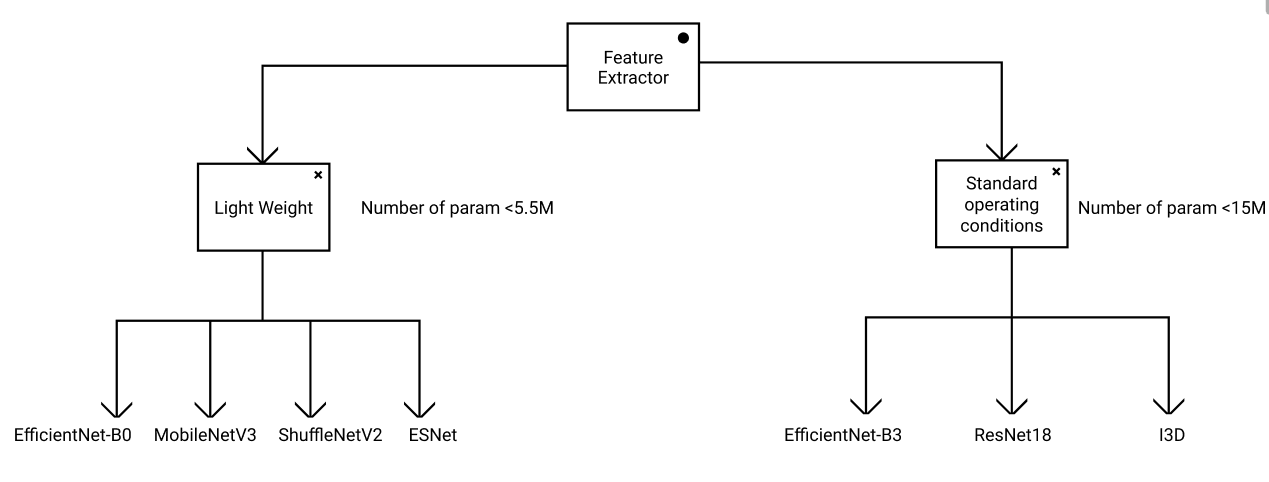
ii) an action recognition model which recognizes autism-related behaviors by modeling the temporal evolution of the extracted frame-wise features.

-> Comparing 7 feature extractors and 4 learning strategies

| Feature Extractor Backbone | Pros | Cons |
| --- | --- | --- |
| EfficientNet | - Computationally efficient compared to ResNet, I3D, etc. - Robust features due to the compound scaling mechanism | - Not suitable for mobile applications. |
| MobileNet | - Computationally efficient. - Does not require an impressive amount of data to train. - Suitable for mobile and embedded vision applications. - Low computational resources are required. | - Limited model capacity to learn large-scale data. - Produce less accurate results compared to EfficientNet, ResNet and I3D |
| ShuffleNet | - Computationally more efficient than MobileNet. - Does not require an impressive amount of data to train. - Suitable for mobile and embedded vision applications. - Low computational resources are required. - More accurate than MobileNet. | - Limited model capacity to learn large scale complex data. |
| ESNet | - Computationally more efficient than MobileNet and ShuffleNet. - Real-time performance. - Suitable for mobile and embedded vision applications. - Does not require an impressive amount of data to train. - Low computational resources are required. | - Limited model capacity to learn large scale complex data. |
| ResNet | - Produce more accurate results than the shallow networks - Suitable for learning large-scale complex data. - The residual connections overcome the vanishing gradient problem. | - Requires more memory and computational resources, which makes it less suitable for mobile applications. |
| I3D | - More suitable for video-based applications to capture spatio-temporal features. - Suitable for learning large-scale complex data. - Produce more accurate results than shallow networks such as MobileNet etc. | - Requires more memory and computational resources, which makes it less suitable for mobile applications. |

-> We select feature extractors appropriate for two different operating conditions:   
 i) a light-weight environment condition suitable for an embedded device

ii) a standard operating condition (i.e. conventional models) suitable for a powerful workstation or server.

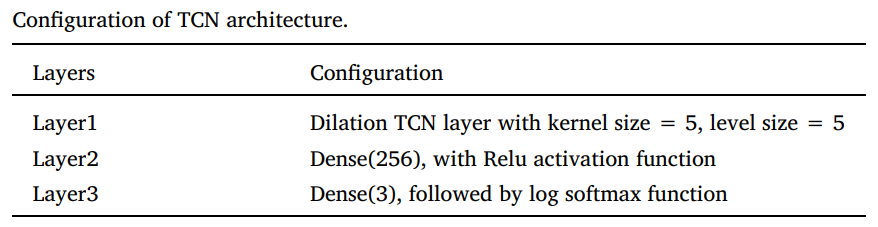


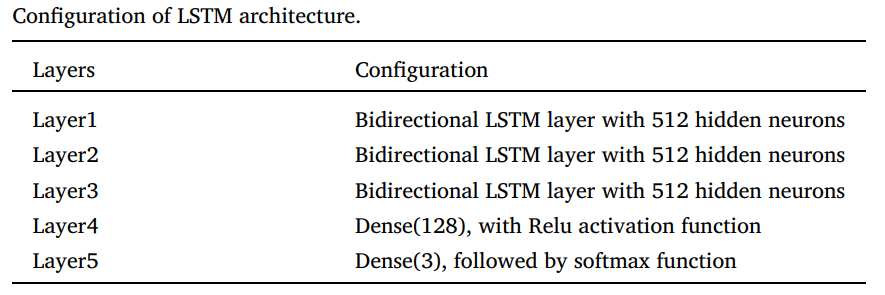
-> more complex networks  
 => ResNet101\_vd 44M

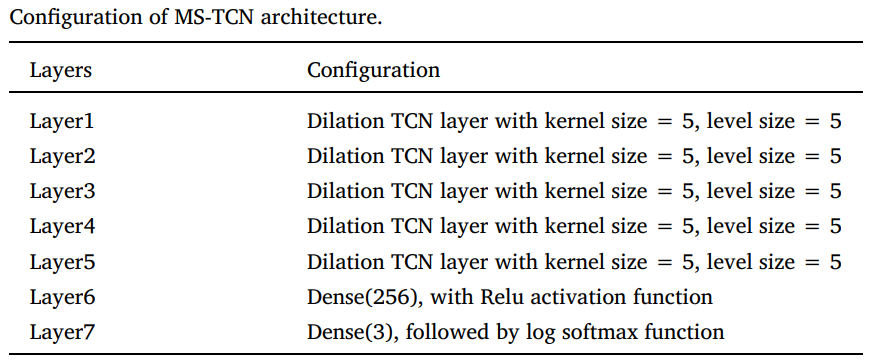
=> HRNet\_W48 78M

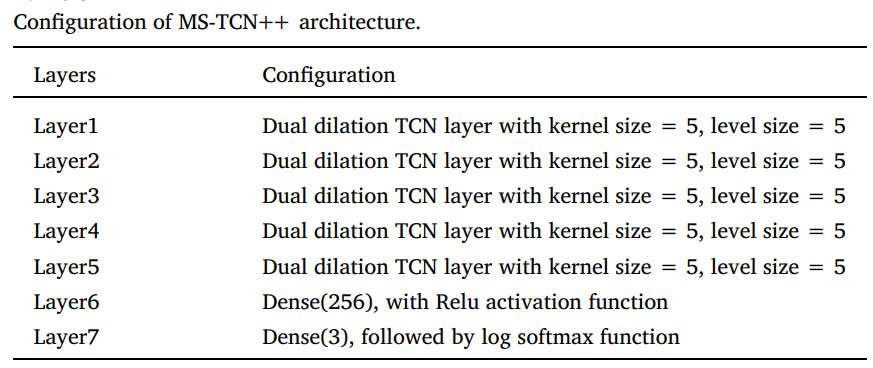
=> SwinTransformer 99M

| Action Recognition Method | Pros | Cons |
| --- | --- | --- |
| LSTM | - Has gates that control the flow of information in and out of the cell. - Low memory requirements. - Faster training. | - May suffer from the vanishing gradients when the data sequence is very long. - May overfit if not properly regularized. |
| TCN | - Able to perform temporal mapping of very long sequences well due to its multiple layers of dilated convolutions (large receptive field). - Can handle variable-length sequences. - Does not suffer from vanishing/exploding gradients. - Able to make more accurate predictions. - Faster training. | - Higher memory requirements compared to LSTM |
| MS TCN | - Achieves better results than TCN as it operates on full temporal resolution (no pooling layers are used). - Able to perform temporal mapping of very long sequences well due to its multiple layers of dilated convolutions (large receptive field). - Can handle variable-length sequences. - Does not suffer from vanishing/exploding gradients. - Multi-stage architecture provides an improved receptive field. - Faster training | - Higher memory requirements compared to LSTM |
| MS TCN++ | - Similar capabilities to MSTCN, however, MSTCN++ achieves better results than MSTCN due to the dual dilated layer. | - Higher memory requirements compared to LSTM |

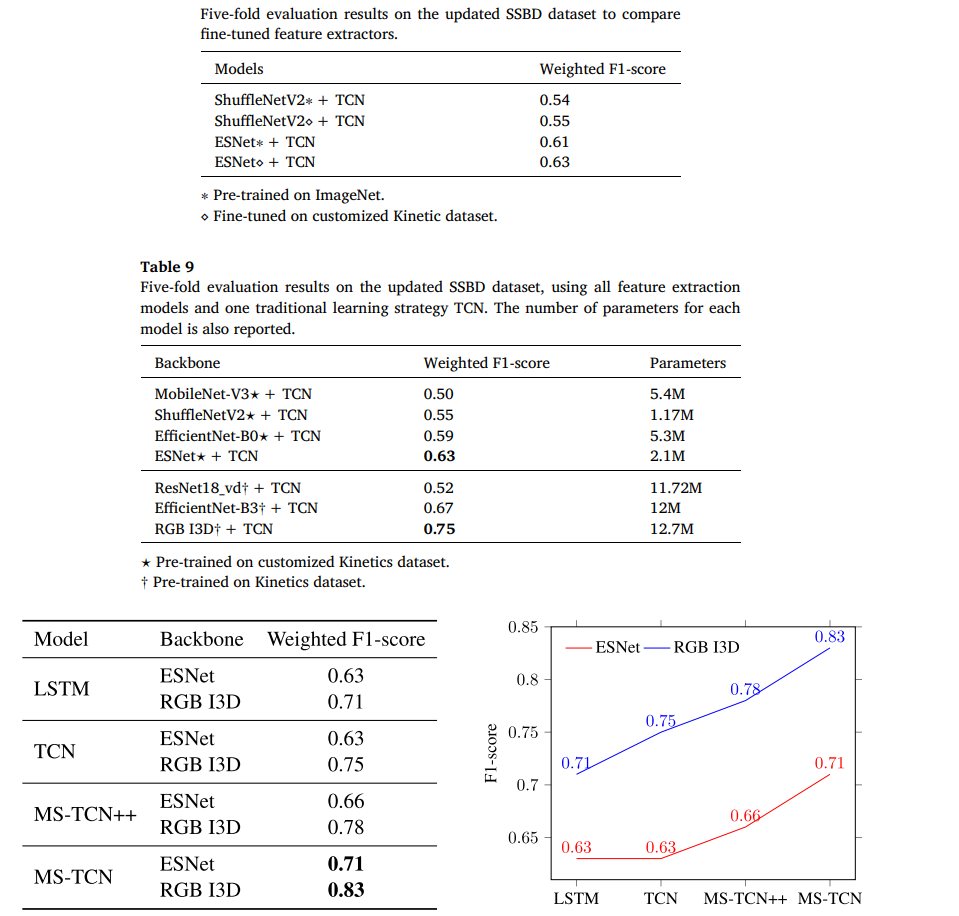








-> result



-> The experimental procedures involving human subjects described in this paper were approved by the CSIRO Health and Medical Human Research Ethics Committee (CHMHREC). The CHMHREC is an NHMRC Registered Human Research Ethics Committee (EC00187). CSIRO Ethics ID 2022\_004\_LR

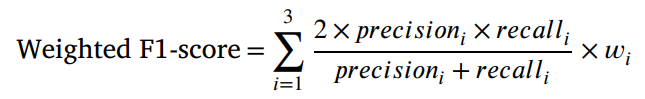
-> The SSBD dataset contains three typical action classes: [arm flapping, headbanging, and spinning](http://refhub.elsevier.com/S2405-8440(23)03970-1/bib4D9D9730B08D8B0C61D9383EE9EEA541s1)

-> pre processing of dataset by cropping the subject alone and removing noisy data from the database

-> pre processing done by a [Mask-RCNN](http://refhub.elsevier.com/S2405-8440(23)03970-1/bib08C68D5A15B6099E3974E116F67B7866s1) pretrained on [COCO](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibC01EB5E415C8DE0E0940FB894AA903ECs1) and implemented through [Detectron2](https://github.com/facebookresearch/detectron2), then resized to a fixed size using the bounding box size

-> 5-fold validation and “Leave-One-Group-Out” setting

-> mean Weighted F1-score for 5-fold validation



-> 200 epochs, batch size=16, learning rate=e^-3, first moment decay rate=0.9, second moment decay rate=0.999

**Activity Recognition with Moving Cameras and Few Training Examples: Applications for Detection of Autism-Related Headbanging**

-> ASD Diagnosis, preprocessed data, collected data, OpenPose realtime multi-person pose estimation, estimated skeletal pose, extracting key points in the head region for accuracy, convolutional neural network (CNN) using the Keras [10] Python library with a Tensorfow [1] backend, a long short-term memory (LSTM) [28] neural network, 3-fold cross validation, 90.77% across the 3 cross-validation folds. The individual F1-scores per fold are 83.3%, 89.0%, and 100.0%.

* -> [“ASD is a neurodevelopmental disorder characterized by a set of social communication deficits, self-harm, or persistent repetition of actions”](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibABB7B79D860AFD1314264C986AF4606Cs1)
* -> “ASD related biomarkers such as those determined through functional magnetic resonance imaging (fMRI). [Facial expressions](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibD2F6BC878DD136CE1116D13D2DF5BCA9s1), [eye gaze, and motor control/movement patterns](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibF685F61EDB75CC4742D43DD621AEA962s1)”
* To minimize overfitting, rotation at a random interval between -45 and 45 degrees and zooming in with a random zoom factor between 1.0 and 2.0.
* dataset from SSBD from [rajagopalan et al](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibA1DD584B1FB446E21CB1B438E6012CB9s1).
* 27 video clips containing headbanging video clips containing “normal” head motions.
* OpenPose realtime multi-person pose estimation to track skeletal keypoints in each frame
* To account for the body part occlusion issue, which would inject unnecessary noise which would confuse the classifer
* extracting key points in the head region.
* We implement a time-distributed convolutional neural network (CNN) using the Keras [10] Python library with a Tensorfow [1] backend.
* We train using Adam optimization [34] with an initial learning rate of 0.0001.
* We perform 3-fold cross validation, ensuring that no child who appeared in the train set would appear in the test set for all folds.
* To minimize overftting and increase generalization, we apply the following data augmentations to each frame: rotation at a random interval between -45 and 45 degrees and zooming in with a random zoom factor between 1.0 and 2.0.

**Pros And Cons of Dense Optical Flow:**

Pros : Dense optical fow computes fow for all points in a frame, resulting in “flow vectors” with a magnitude and direction.

Cons : extra non-relevant movement patterns add much noise to the dense optical fow image, making it a non-ideal representation.

**Pros And Cons of Lucas Kanade Optical Flow:**

Pros : Lucas-Kanade optical fow, in contrast to dense optical fow, computes fow for a sparse number of points pre-defned by the user [3], for example detected edges or corners.

Cons:The Lucas-Kanade method brings its own set of limitations by detecting movement outside of the body to an extent that is more dramatic than dense optical fow. This technique is particularly sensitive to background movement detection associated with slight shifts in the camera.

**Pros And Cons of Pose Estimation:**

Pros: Pose estimation is a technique which is more robust to camera movement compared to optical flow. . OpenPose uses a CNN which is trained to predict part afnity felds, or fow felds representing relationships between body parts, and confdence maps which encode body part locations. Unlike optical fow, OpenPose predicts each frame independently of the surrounding frames.

Cons: There are some clear limitations to using unmodifed pose estimation. The noisy skeleton problem is a documented issue which arises when body parts are self-occluded. We observe that body part occlusion is a frequent occurrence in unstructured home videos, making unmodifed pose estimation a non-ideal feature representation.

**Application of Skeleton Data and Long Short-Term Memory in Action Recognition of Children with Autism Spectrum Disorder**

-> asd diagnosis, skeletal data + LSTM

-> OpenPose algorithm for skeletal data

-> 4 denoising techniques applied to skeletal data before passing to stage 2

-> stage 2 = tracking humanoids based on temporal data

-> stage 3 = LSTM classification

-> OpenPose + denoising > tracking > classification

-> [characterized by persistent deficits in social communication and interaction as well as restricted and repetitive behaviors](https://sci-hub.live/10.1016/b978-0-12-809324-5.05530-9)

-> First, we scale the coordinates of the key points outputted by OpenPose to the same units

-> Second, we remove the five joints on the head

-> Third, we discard frames without skeleton data or missing important joints.

-> Finally, we use the relative joint positions in adjacent frames to fill in the unrecognized joint positions

-> head joints provides less information on the activity classification, thus they removed the head joints from the maps

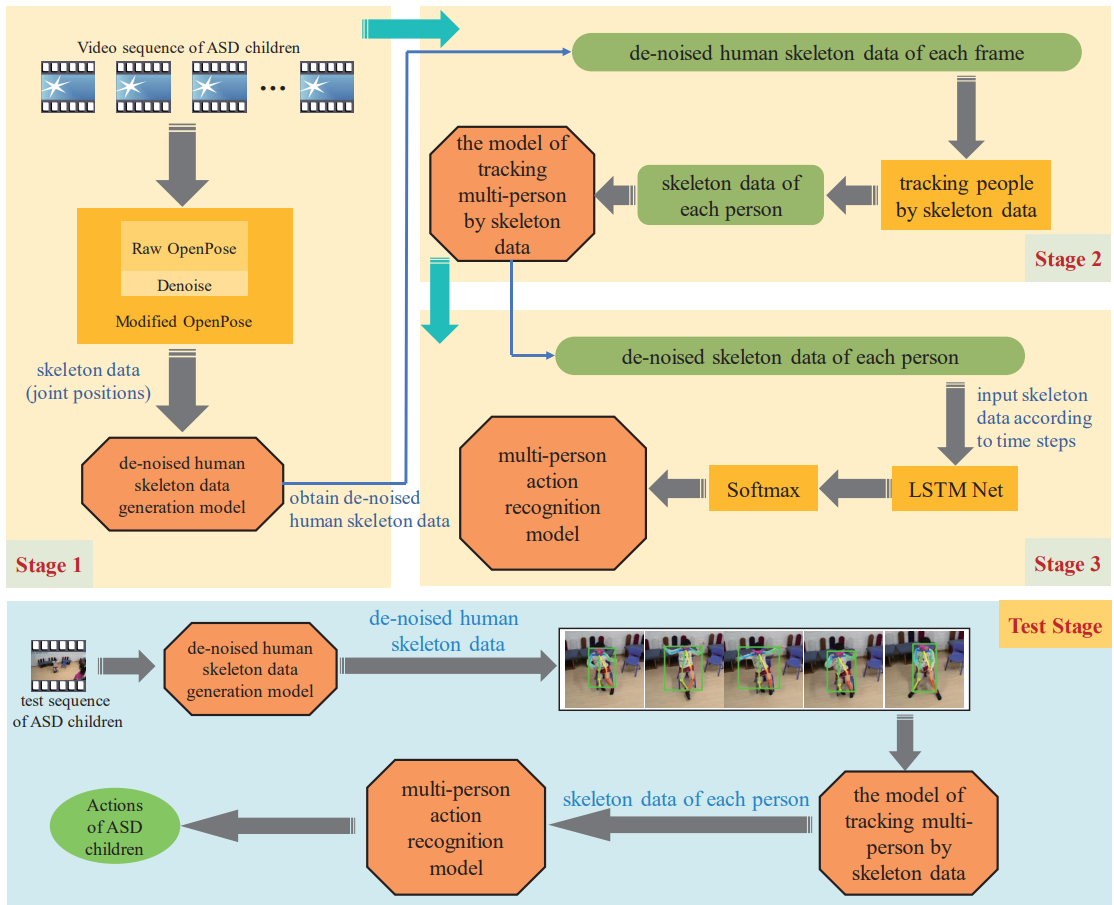
-> frames with major joints missing will be discarded

-> other joints missing will be filled in by the temporal frames adjacent to them

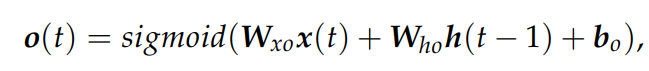
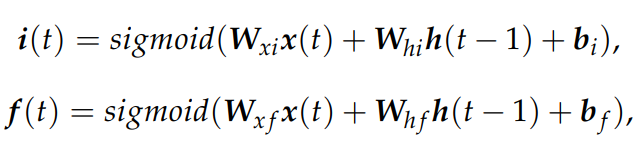
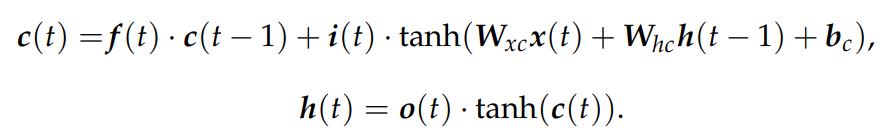
-> facilitates tracking of multiple skeletons

-> DeepSORT and SORT are computationally expensive

-> a new algorithm is developed which finds the distance of the joint i of the jth person id and calculates the relative position to the frame’s center, this location is compared with adjacent temporal frames to identify if it is the same person for tracking the skeleton





-> collected data, 5 actions, sit stand squat shake body shake hands, 1062 sequences

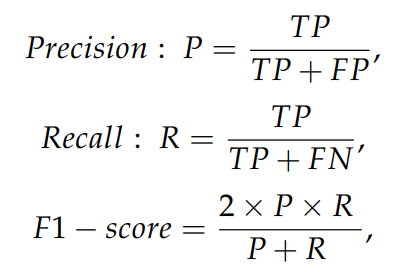
-> normalized to 656x368 -> sent to OpenPose

-> for training the model, only one subject should taken in case of multiple subjects in frame, the subject closest to the center of the frame was taken

-> 70-30 split, 32x13x2, 32 frames, 13 skeletal key points, 2 x-y coordinates

-> hidden states = 128, learning rate =0.001, time steps =32, prob\_threshold=0.9

-> mentioned [AlphaPose](https://ieeexplore.ieee.org/abstract/document/9954214/) is also a good method to implement



-> Compared other methods

-> DNN, SVM, decision tree, RF

-> F1 score is used for performance metric

Action Original De-noised

Sit 0.734 0.8868

Stand 0.9279 0.9896

Squat 0.7757 0.8925

Shake body 0.8604 0.9415

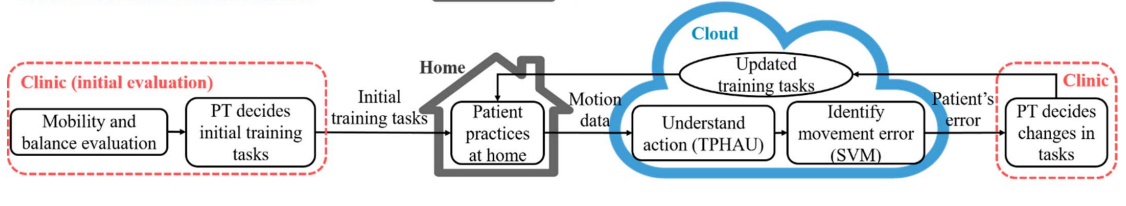
Shake hand 0.9094 0.9803

**Human Action Understanding and Movement Error Identification for the Treatment of Patients with Parkinson’s Disease**

-> Parkinson’s disease (PD) is one of the most common neurodegenerative movement disorders, especially in the elderly. [More than 10 million people worldwide are living with PD. In the US, about 60,000 people are diagnosed with PD each year](https://parkinsonsnewstoday.com/parkinsons-disease-statistics/). The major motor symptoms of PD include tremor, rigidity, and postural instability.

-> learning-based personalized treatment system to enable home-based training for PD patients.

-> uses the **Kinect sensor** for input, a *two-phase human action understanding* algorithm **TPHAU** to understand the patient’s movements, **Support Vector Machine** to evaluate patient performance.

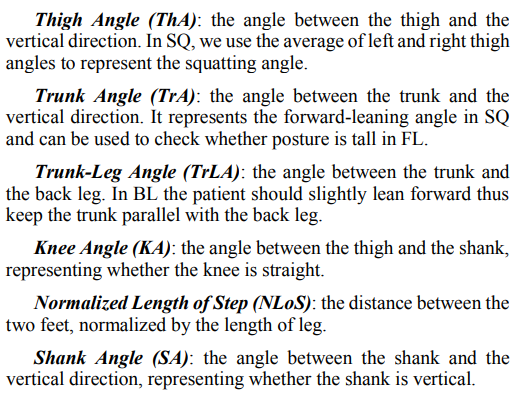


-> **HMM(Hidden Markov model)** - used in **TPHAU**

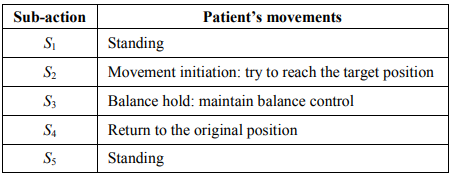
-> Tasks assigned by PT : squat (SQ), forward lunge (FL) and backward lunge (BL). Each varying from level 1 to 4.

-> **KCQs** : Kinect-captured quantities.

-> Kinect sensor captures 25 joints of the human skeleton with 3-D coordinates for each joint.**KCQs** are quantities that can be derived from the joint coordinates captured by Kinect. In this paper, we define the following six KCQs for the three tasks :



-> Algorithm based on **HMM** to detect sub-actions(sub-actions makes segmentation difficult) in patient’s movements.

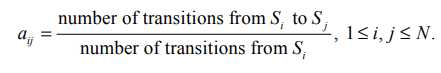


-> **HMM** : a statistical Markov model that assumes the system to be a Markov process with hidden states.

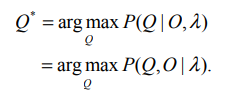
-> HMM model characterized by : hidden state set, HMM feature, observation sequence, hidden state sequence, state transition probability matrix, emission probability.

-> Our goal is to infer the hidden state sequence Q from the observation sequence O, then segment the interval of each sub-action according to the hidden state sequence. (There's no need to classify a new movement into an optimal task as the task that the patient is performing is known)

-> The transition probability matrix is calculated as :

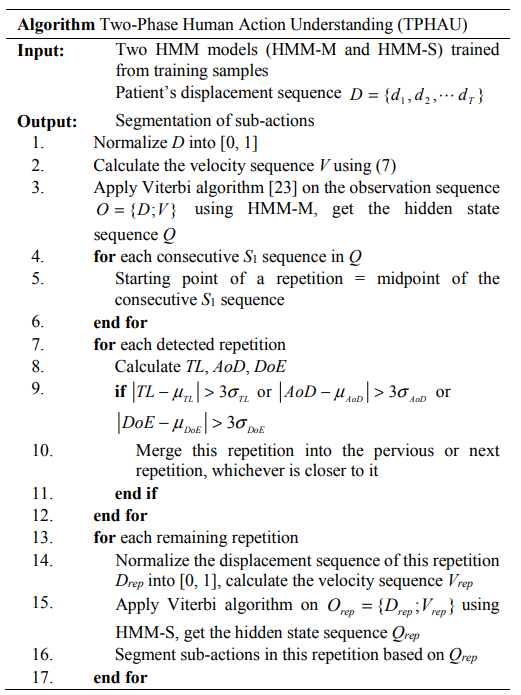


-> [The Viterbi algorithm](http://dx.doi.org/10.1109/PROC.1973.9030) is a dynamic programming algorithm for finding the most likely hidden state sequence Q\* of the observation O by :



-> **TPHAU Algorithm :**

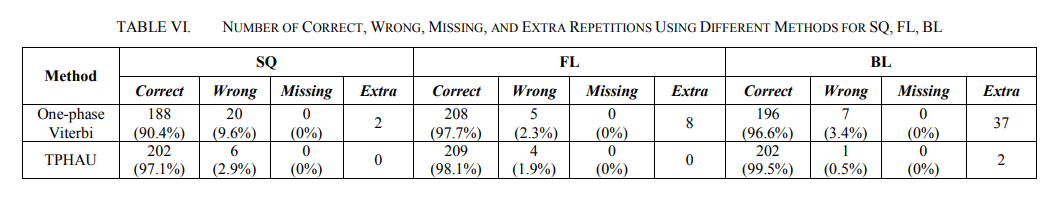
Proposed based on the selected features (d, v) and the [Viterbi algorithm](http://dx.doi.org/10.1109/PROC.1973.9030) that is used to infer the optimal hidden state sequence from the observation.

****

-> It detects patient’s repetitions in performing the task in the first phase, and segments sub-actions in each repetition in the second phase.

-> noise may cause the detection of extra repetitions. There are mainly two types of extra repetitions

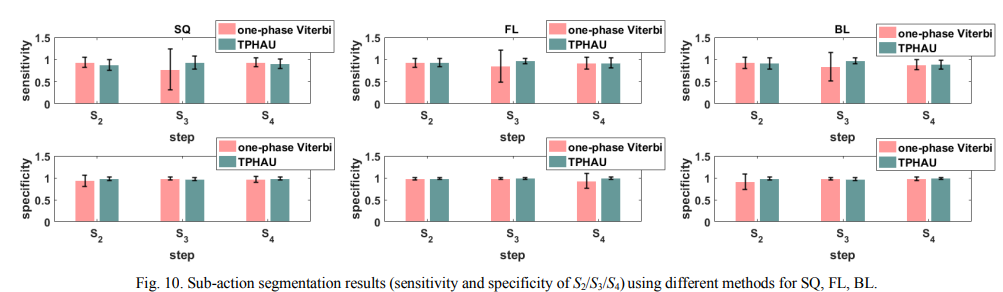
1. Noise being detected as complete repetitions
2. Recognizing one repetition as two or more

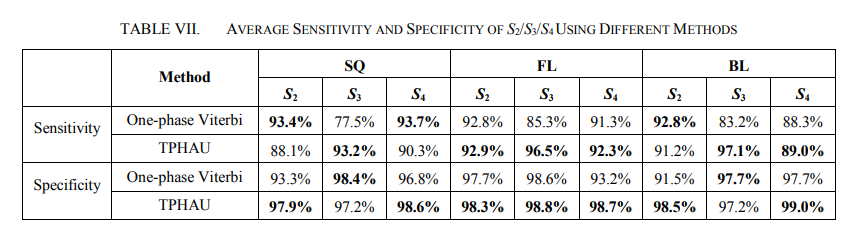


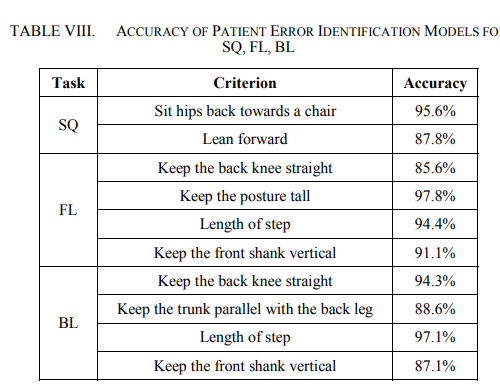
**Accuracy :** The average and 95% confidence interval are calculated among all the test samples.

**Criterion wise eval :** For each criterion, the accuracy is calculated as the ratio of the correctly classified samples to the total number of test samples.

=>accuracy of each criterion for the three tasks are above 85%







-> SVM-based method to identify patient’s error in performing the task, which emulates PT evaluation

-> The proposed treatment system enables low-cost and remote care for PD patients.

**->DATASET :**

* The dataset used in the study consists of motion data collected from 26 Parkinson's disease (PD) patients, with ages ranging from 56 to 89, including 15 males and 11 females..
* The motion data was collected using the Microsoft Kinect sensor, capturing the (x, y, z) coordinates of 25 joints per frame, resulting in approximately 90,000 data points for each task performed by a patient in one session.
* The dataset includes samples for different tasks, such as Sit-to-Stand (SQ), Forward Lunge (FL), and Backward Lunge (BL), with variations in difficulty levels based on the patient's health condition

**-> PROS AND CONS OF HMM :**

| **PROS** | **CONS** |
| --- | --- |
| HMM can effectively model temporal sequences and is suitable for capturing the dynamics of patient motion data. | HMM may require training and tuning for specific applications, which can be time-consuming. |
| it can be used for tasks such as repetition detection and sub-action segmentation within each repetition. | It may not be suitable for handling complex training tasks and criteria that apply to the entire task, especially for patients with more complex balance and agility programs. |
| HMM can compensate for the temporal variation of patient's motion data, making it suitable for evaluating patient's performance accurately. |  |

**-> PROS AND CONS OF TPHAU:**

| PROS | CONS |
| --- | --- |
| TPHAU addresses the delay problem and evaluates patient's performance accurately. | The implementation of TPHAU may require additional computational resources and complexity compared to simpler models. |
| It can detect patient's repetitions in performing tasks in the first phase and segment sub-actions in each repetition in the second phase. | TPHAU may require careful parameter tuning and validation to ensure optimal performance. |
| TPHAU can enhance the accuracy in sub-action detection and improve the specificity and sensitivity of the results. |  |

**HuGaDB: Human Gait Database for Activity Recognition from Wearable Inertial Sensor Networks**

**GAIT ANALYSIS :** Gait analysis focuses on recognizing activities and how they are performed. It can be used in healthcare systems for monitoring patients recovering after surgery, fall detection, or diagnosing conditions like Parkinson’s disease.

-> The **HuGaDB** dataset provides detailed kinematic data for analyzing human gait and activity recognition, collected from 18 healthy participants. -> This dataset can be used in health-care-related studies, such as walking rehabilitation, or in modeling human movements in virtual reality or humanoid robotics.

-> a human gait data collection for analysis and activity recognition consisting of continuous recordings of combined activities, the data recorded are segmented and annotated.

-> data collected from a body sensor network consisting of 6 wearable sensors(accelerometer and gyroscope) : right and left thighs, shins, and feet

-> two electromyography sensors were used on the quadriceps (front thigh)8

->Data can be loaded easily in most popular programming languages

-> Sample data with respect to the activities are visualized through a heat map representation

-> Data are scaled to the range.

| **Human activity recognition (HAR)** | **Gesture recognition (GR)** |
| --- | --- |
| Human activity recognition (HAR) aims to recognize daily lifestyle activities, such as cooking, cleaning, or walking. It involves on-body inertial sensors and additional sensors like temperature, proximity, and heart rate sensors. Databases for HAR include MIT Place dataset, Darmstadt Daily Routine dataset, and others. | On the other hand, gesture recognition (GR) mainly focuses on recognizing hand-drawn gestures in the air, such as numbers, circles, or letters. It typically uses smartphone sensors or special gloves equipped with kinematic sensors and electromyography sensors. The databases for gesture recognition are distinct from those used for HAR. |

**DATABASE LINK :** [click](https://github.com/romanchereshnev/HuGaDB)

**A Feasibility Study of Autism Behavioral Markers in Spontaneous Facial, Visual, and Hand Movement Response Data**

**ASD :** Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects multiple areas of an individual’s verbal and nonverbal communication skills to varying extents. The onset of ASD occurs during the early developmental period of the child with the presentation of subtle deviances from typical behavioral and cognitive traits.

-> Commonly identified through direct observation of atypical behaviors.

-> measurement tool known as the Broader Phenotype Autism Symptom Scale [(BPASS)](https://link.springer.com/article/10.1007/s10803-006-0182-2)

-> Although effective, the BPASS method entails significant time, cost, training, and expertise due to human interventions

->

**Facial Signs and Psycho-physical Status Estimation for Well-being Assessment**

* Stress is a state being present as a part of pressures of accelerated life rhythms.
* Anxiety is, in general terms, the unpleasant feeling of worrying, fear and uneasiness when a perceived threat is present.
* Under certain circumstances, anxiety can be conceived as a mental disorder called generalized anxiety disorder, characterized by a disproportionate uncontrollable and irrational worry in common life

Activities.

* Facial Signs of stress - gaze spatial distribution, saccadic eye movement, pupil diameter increase, high blink rates, jaw clenching, grinding teeth, trembling of lips, and frequent blushing
* Facial Signs of anxiety - reddening, lip deformations, eye blinking, strained face, facial pallor, dilated pupils, and eyelid twitching

-> **Facial Parameter Representation Systems**

* **Facial Action Coding Systems (FACs) :**

FACS is a comprehensive system, standardizing sets of muscle movements known as action units (AUs) that produce facial expressions. It consists of 44 unique AUs and their combinations, each of which reflects distinct momentary changes in facial appearance.

However, due to its subjective nature, FACS may be biased and image annotation is time consuming

* **MPEG-4 with Facial Animation Parameters (FAPs) :**

The MPEG-4 standard supports facial animation by providing 66 low-level “FAPs”. The FAPs represent a complete set of facial actions, along with head motion, tongue, eye and mouth control, which deform a face model from its neutral state. The FAP value indicates the magnitude of the deformation caused on the model.

->**Facial Feature Extraction Methods**

* **Muscle based Methods :**

Muscle/geometric based techniques of feature extraction estimate facial muscle deformation from videos and images. A facial expression is a result of one or more facial features due to the contraction of the muscles of the face. Facial features change either

their motion or position (eye, eyebrows, nose and mouth) or their geometric characteristics and shape.

In the FAP extraction process the outer-lip and eyebrow contours are tracked for each frame and compared to the corresponding contours of the neutral frame of the sequence in order to calculate FAPs in terms of facial animation parameter units.

* **Model based Methods :**

Model (appearance) based methods depend on the general appearance of the face or/and specific regions (skin texture, wrinkles, furrows etc.) for fitting 2D or 3D face models.

Active appearance models (AAMs) provide a consistent representation of the shape and appearance of the face and have been used for facial expression classification

* **Motion based Methods :**

Motion based methods utilize features derived from movements, extracted from image sequences of either facial components or the whole face.

Optical flow is the method adopted by the majority of facial motion analysis systems.

Facial action coding system (FACS) and FACS+ is also used

* **Hybrid Methods :**

Hybrid methods use features from local and holistic approaches in order to improve the results. These approaches reflect the human perception that utilizes both local face signs and the whole face to recognize a facial expression.

-> **Facial Expression Coding**

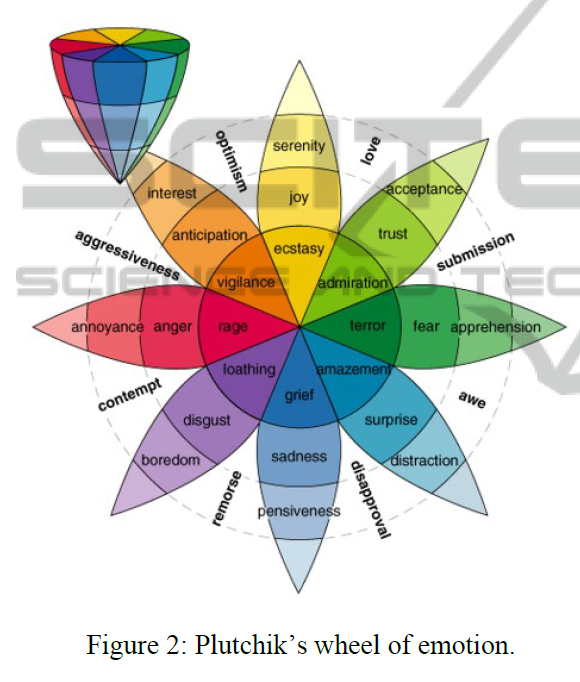
* Principal component analysis (PCA), linear discrimination analysis (LDA) and independent component analysis (ICA) are used for dimensionality reduction of facial features.

-> **Classification of Facial features**

* Supervised learning methods - hidden Markov models, dynamic Bayesian networks and naive Bayes classifiers.
* Unsupervised learning methods - geometric-invariant clustering, aligned cluster analysis and AAMs
* the main kinds of approaches for converting the extracted features into a facial expression class are machine learning classifiers and rule-based systems
* The main machine learning classifiers can be linear or multiclass and include SVMs, AdaBoost or GentleBoost, k-nearest neighbours, multivariate logistic regression and multilayer neural networks

-> **Coding Classes into Psycho-physical Statuses**

* A conception of emotions, called the “wheel of emotions”, demonstrates how different emotions can blend into one another and create new emotions



* Action units can be divided into primary and auxiliary. Their combination formulates a facial expression.

**Automated Detection Approaches to Autism Spectrum Disorder Based on Human Activity Analysis: A Review**

**Datasets Used:**

Skeleton Dataset : Skeletal data encode the human body posture in terms of relative 3D coordinates of different body joints and the joints’ orientation angle. Rihawi et al.

Rihawi O, Merad D, Damoiseaux JL. 3D-AD: 3D-autism dataset for repetitive behaviours with kinect sensor. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE; 2017. p. 1–6.

Video Dataset : To the best of our knowledge, Zunino et al. [1] developed the only available video dataset to analyze the action style of an autistic child. The dataset includes activities such as the task of placing, picking, passing, and grasping a bottle of a particular size by an autistic child.

Zunino A, Morerio P, Cavallo A, Ansuini C, Podda J, Battaglia F, etal. Video gesture analysis for autism spectrum disorder detection. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE; 2018. p. 3421–3426.

Gaze Dataset : Eye-tracking refers to the process of measuring visual attention. These measurements are captured using an eyetracking device that records the positions and movements that our eye makes while viewing a scene. To this end, Duan et al. [4] developed an eye movement dataset named ‘Saliency4ASD’ from 14 ASD and 14 typically developed children.

Duan H, Zhai G, Min X, Che Z, Fang Y, Yang X, et al. A dataset of eye movements for the children with autism spectrum disorder. In: Proceedings of the 10th ACM Multimedia Systems Conference; 2019. p. 255–260.

Yaneva et al. [5] also developed a gaze dataset, which included autistic adults instead of children.

Yaneva V, Ha LA, Eraslan S, Yesilada Y, Mitkov R. Detecting Autism Based on Eye-Tracking Data from Web Searching Tasks. In: Proceedings of the Internet of Accessible Things. W4A ’18. New York, NY, USA: Association for Computing Machinery; 2018. p. 1–10.

In recent work, Shihab et al. [51] provided a gaze dataset that comprises face-scanning data of adults and children diagnosed with ASD.

Shihab AI, Dawood FA, Kashmar AH. Data Analysis and Classifcation of Autism Spectrum Disorder Using Principal Component Analysis. Adv Bioinform. 2020;2020:1–8.